
Table of Contents

Foreword.....	vii
Preface.....	ix
1. Analyzing Big Data.....	1
The Challenges of Data Science	3
Introducing Apache Spark	4
About This Book	6
2. Introduction to Data Analysis with Scala and Spark.....	9
Scala for Data Scientists	10
The Spark Programming Model	11
Record Linkage	11
Getting Started: The Spark Shell and SparkContext	13
Bringing Data from the Cluster to the Client	18
Shipping Code from the Client to the Cluster	22
Structuring Data with Tuples and Case Classes	23
Aggregations	28
Creating Histograms	29
Summary Statistics for Continuous Variables	30
Creating Reusable Code for Computing Summary Statistics	31
Simple Variable Selection and Scoring	36
Where to Go from Here	37
3. Recommending Music and the Audioscrobbler Data Set.....	39
Data Set	40
The Alternating Least Squares Recommender Algorithm	41
Preparing the Data	44

Building a First Model	46
Spot Checking Recommendations	48
Evaluating Recommendation Quality	50
Computing AUC	51
Hyperparameter Selection	53
Making Recommendations	55
Where to Go from Here	56
4. Predicting Forest Cover with Decision Trees.	59
Fast Forward to Regression	59
Vectors and Features	60
Training Examples	61
Decision Trees and Forests	62
Covtype Data Set	65
Preparing the Data	66
A First Decision Tree	67
Decision Tree Hyperparameters	71
Tuning Decision Trees	73
Categorical Features Revisited	75
Random Decision Forests	77
Making Predictions	79
Where to Go from Here	79
5. Anomaly Detection in Network Traffic with K-means Clustering.	81
Anomaly Detection	82
K-means Clustering	82
Network Intrusion	83
KDD Cup 1999 Data Set	84
A First Take on Clustering	85
Choosing k	87
Visualization in R	90
Feature Normalization	91
Categorical Variables	94
Using Labels with Entropy	95
Clustering in Action	96
Where to Go from Here	97
6. Understanding Wikipedia with Latent Semantic Analysis.	99
The Term-Document Matrix	100
Getting the Data	102
Parsing and Preparing the Data	102
Lemmatization	104

Computing the TF-IDFs	105
Singular Value Decomposition	107
Finding Important Concepts	109
Querying and Scoring with the Low-Dimensional Representation	112
Term-Term Relevance	113
Document-Document Relevance	115
Term-Document Relevance	116
Multiple-Term Queries	117
Where to Go from Here	119
7. Analyzing Co-occurrence Networks with GraphX.....	121
The MEDLINE Citation Index: A Network Analysis	122
Getting the Data	123
Parsing XML Documents with Scala's XML Library	125
Analyzing the MeSH Major Topics and Their Co-occurrences	127
Constructing a Co-occurrence Network with GraphX	129
Understanding the Structure of Networks	132
Connected Components	132
Degree Distribution	135
Filtering Out Noisy Edges	138
Processing EdgeTriplets	139
Analyzing the Filtered Graph	140
Small-World Networks	142
Cliques and Clustering Coefficients	143
Computing Average Path Length with Pregel	144
Where to Go from Here	149
8. Geospatial and Temporal Data Analysis on the New York City Taxi Trip Data.....	151
Getting the Data	152
Working with Temporal and Geospatial Data in Spark	153
Temporal Data with JodaTime and NScalaTime	153
Geospatial Data with the Esri Geometry API and Spray	155
Exploring the Esri Geometry API	155
Intro to GeoJSON	157
Preparing the New York City Taxi Trip Data	159
Handling Invalid Records at Scale	160
Geospatial Analysis	164
Sessionization in Spark	167
Building Sessions: Secondary Sorts in Spark	168
Where to Go from Here	171

9. Estimating Financial Risk through Monte Carlo Simulation.....	173
Terminology	174
Methods for Calculating VaR	175
Variance-Covariance	175
Historical Simulation	175
Monte Carlo Simulation	175
Our Model	176
Getting the Data	177
Preprocessing	178
Determining the Factor Weights	181
Sampling	183
The Multivariate Normal Distribution	185
Running the Trials	186
Visualizing the Distribution of Returns	189
Evaluating Our Results	190
Where to Go from Here	192
10. Analyzing Genomics Data and the BDG Project.....	195
Decoupling Storage from Modeling	196
Ingesting Genomics Data with the ADAM CLI	198
Parquet Format and Columnar Storage	204
Predicting Transcription Factor Binding Sites from ENCODE Data	206
Querying Genotypes from the 1000 Genomes Project	213
Where to Go from Here	214
11. Analyzing Neuroimaging Data with PySpark and Thunder.....	217
Overview of PySpark	218
PySpark Internals	219
Overview and Installation of the Thunder Library	221
Loading Data with Thunder	222
Thunder Core Data Types	229
Categorizing Neuron Types with Thunder	231
Where to Go from Here	236
A. Deeper into Spark.....	237
B. Upcoming MLlib Pipelines API.....	247
Index.....	253